Network Streaming and Compression for Mixed Reality Tele-Immersion

Mekuria, R.N.

2017

document version
Publisher's PDF, also known as Version of record

Contents

Chapter 1 Introduction

Humans enjoy communicating and sharing experiences with each other. The Internet is an excellent platform to facilitate this need. The Internet enables distributed shared experiences such as video conferencing, voice calls (possibly in a group), chatting, online gaming and virtual reality. This is changing the way we interact with each other in our daily lives. Current rapid advances in 3D depth acquisition are enabling near-instant highly realistic 3D capture of real humans and objects. The integration of these 3D representations in distributed shared experiences such as virtual reality, social networking, tele-conferencing and gaming could result in improved and novel user experiences. One example of a novel user experience would be the tele-immersive scenario, where a real users can enter a virtual world and interact with a realistic representation of himself. In this example he or she can interact with computer controlled avatars in a virtual world and with other users. Another example of an improved user experience would be interactive free viewpoint rendering of 3D captured content in a social network portal or photo sharing site.

This thesis will look at the implementation of end-to-end multimedia systems based on such highly realistic 3D representations. It will focus on real-time end-to-end conversational style communications. A typical end-to-end media communications pipeline for 3D data capture is show in Figure 1. 3D Media is captured using 3D depth sensing devices such as motion sensors, microphone arrays and 3D cameras such as Microsoft Kinect. A subsequent reconstruction stage can generate complete 3D representations such as a 3D Point Cloud or a 3D Mesh using a 3D reconstruction algorithm. In video conferencing, typically, compressed 2D video is sent over the communication link. For advanced 3D communication systems based on 3D representations meshes and point clouds have to be compressed and transmitted instead. This poses a large challenge in designing such end-end 3D communication systems. While the compression of 3D data such as point clouds and meshes has been considered in the 3D graphics literature [1], real-time compression of time-varying data for virtual reality like applications has rarely been considered. Nevertheless, this is the type of compression that is needed to enable mixing real and virtual reality in real-time distributed systems.

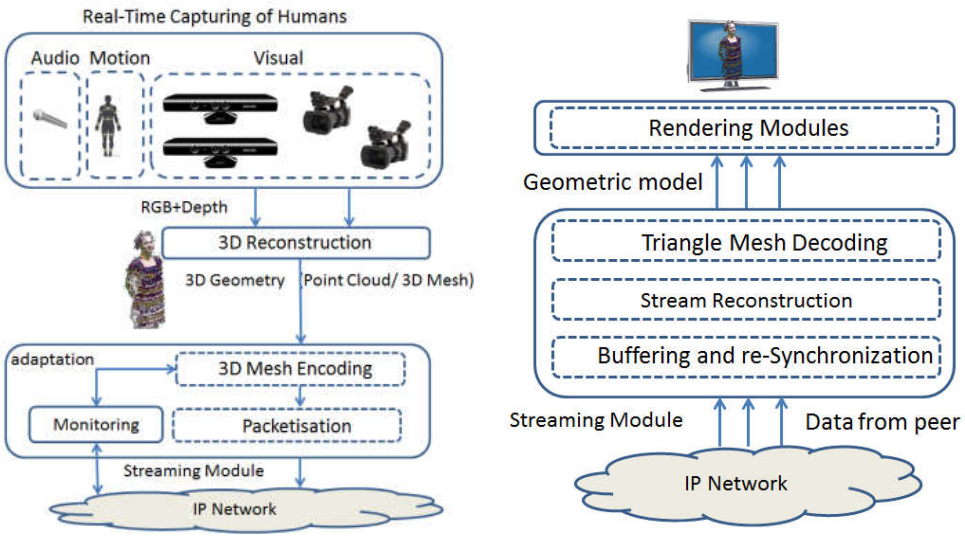

Figure 1 3D Tele-immersive media pipeline

In addition, to support video conferencing style communications technologies for transmission, packetisation, monitoring and media synchronization become important.

The compression is quite challenging due to the unstructured nature of the data and the real-time requirements. The transmission and synchronization are challenging due to the large frame sizes, varying frame rates and frame size variations. This thesis develops three different approaches for the real-time compression of 3D meshes in such a communication scenario. Further, a codec for time varying 3D point clouds is developed, an alternative useful representation for 3D objects. These codecs are integrated in a 3D streaming framework. This framework introduces session management and signalling and a media synchronization strategy suitable for this type of data. The complete framework is then used to experiment in a realistic integrated 3D immersive mixed reality communication system with state of art rendering, 3D data capture and social network services. The research work done in this thesis on end-to-end multimedia systems presents new insights in systems integration, codec design, quality assessment, user experience assessment and media standardisation.

1.1 Motivation

3D Data capture: More 3D devices are becoming available to consumers that include depth sensing and sometimes 3D reconstruction. 3D capture could be based on cameras such as Microsoft Kinect (v1, v2) [2] or based on mobile devices [3] [4]. The combination of these capabilities with computing power enable acquisition of fuller 3D Media representations such as 3D geometry with attributes based on 3D Meshes and/or Point Clouds.

Standardized *Compression Technologies* for media data types are becoming available on the market. Media compression standards often result from joint efforts of industry and academia. Popular standards include MPEG-2 video for digital television, MPEG-2 part 3 for music distribution in the internet (MP3), and the JPEG codec for image compression. For live captured 3D point cloud and mesh data an equivalent widely adopted codec does not exist yet. This motivates us to work on the development of technologies to pursue this ultimate standardisation goal of a 3D Point cloud and mesh compression standard suitable for mass usage in the digital ecosystem.

The *bandwidth and QoS* in access, core and local networks is increasing for mobile and fixed networks. These increased bandwidths will enable fast transmission of 3D data, which can enable distributed shared 3D experiences. In addition, new technologies such as software defined networking can provide per flow bandwidth, loss and jitter constrained communications useful for audio-visual data types. While a lot of work is still to be done in this area, the increasing power of the network and emergence of QoS is a key motivation for our work.

3D Rendering: Real-Time rendering of meshes and point clouds is becoming generically available via API's like Direct 3D, OpenGL and on different software and hardware platforms. The ubiquitous availability of 3D rendering of such 3D primitives makes it easy for consumer to view and interact with this data, even in web based applications.

Social networking : Applications like Facebook, Linkedin and other social network portals make it natural from people from all over the world to interact and share

experiences through video conferencing or photo sharing. Taking this existing interconnectedness of friends, family acquaintances and co-workers into account, a logical next step would be to enable 3D mixed reality in such social networking context. This could be useful in both professional settings (job interview, collaborative work) or in a social setting (social hangout, social gaming). Therefore, the interconnectedness of social networking is a key motivation for our work.

1.2 Challenges

3D Rendering and capturing combined with high bandwidth QoS enabled networking are paving the way for 3D tele-presence in the Internet. Nevertheless, the design, deployment and implementation of such systems still faces significant challenges.

3D Compression: realistic 3D mixed reality uses object based 3D video formats. These are currently not well supported in the popular media codecs such as the ones developed by the Moving Picture Experts Group (MPEG) [5] or Joint Picture Experts Group (JPEG) [6]. Reconstructed 3D Meshes and point clouds often consist of huge amounts of points and triangles consuming large amounts of data. A large part of this thesis will be devoted to efficiently compress these formats. These attempts not only aim at efficient compression of 3D frames, but also at *fast* compression to enable real-time end-to-end interactive communication.

3D Data Transmission: Compressed 3D frames resulting from 3D reconstruction are often large in size (over 1 MB per frame) and lack the data loss resilience such as present in popular image/video codecs. Therefore, a small data loss may result in a complete discard of the frame. In addition, some of our experiments have shown that end-to-end network delay can be problematic when using the TCP protocol to rapidly send large frames. Therefore, for 3D mixed reality communication it is important to study error and delay resilient transmission techniques that can handle large frames in real time. In addition to this, frames may be produced at varying framerates (5-30 fps) and frames may contain different numbers of points. These differences compared to normal video/audio require specific attention to the end-to-end pipeline. The varying frame rates and frame sizes make intra and inter-stream synchronization challenging. In addition, for mixed reality, inter-sender synchronization of frames is important to avoid inconsistence when rendering in the 3D virtual

world. This thesis deals with these different challenges in 3D data transmission and synchronization in the context of a practical system.

Signalling and Session Management: Many different types of 3D data exist such as animations, textures, meshes, point clouds and spatial audio. They need to be supported by appropriate signalling and session management. In addition, the architecture for 3D mixed reality envisioned in this thesis can also support different types of users connected to a social network. Light clients, with rendering only capabilities only up to heavy clients with multi camera reconstruction will be supported. Between users present in the system, appropriate media streams need to be setup for communication. The variety of modules, local configurations and different bit-rate codec settings make signalling and setup of the media sessions a challenging problem. This is not handled by existing media protocols. This thesis develops a signalling platform for highly heterogeneous clients using different types of 3D data formats. While a first attempt, it aims to uncover some of the challenges related to this aspect of 3D immersive communications.

System Integration: 3D mixed reality systems are complex systems. They combine 3D capturing technologies, 3D rendering, networking and Artificial intelligence. This highly interdisciplinary mix of computer software creates problems and challenges of its own. Each of the processes can run in their own threads, possibly consuming a disproportional amount of resources. In addition, they often use different libraries that possibly result in clashes (incompatibility, namespace, Operating systems issues). Last the data flow between modules needs to be carefully considered to avoid problems related to processes running in different threads and information inter dependencies.

Quality Assessment: the inter connect of different capture, rendering, compression and transmission methods for novel data types in novel applications introduces the question how to assess quality? Traditional methodology developed for image and video quality assessment cannot be applied directly to this scenario. Different applications, rendering techniques, data types make quality assessment challenging.

1.3 Large Scale Tele-Immersive Systems Architecture

Figure 2 introduces a large scale systems architecture for 3D Tele-immersive communications in a virtual world. The architecture was developed in the REVERIE FP7 project [7] and has driven both the research questions and many of the practical efforts. The architecture includes a social network that enables social network users to join 3D Tele-immersive sessions. The rest of this architecture features components for 3D Tele-immersive communications divided in 4 groups of functionalities.

The yellow *3D capture components* are used for 3D Tele-immersive data capture and reconstruction. These components deal with extracting meaningful 3D information out of visual and aural data streams. These include 3D visual reconstruction, 3D user analysis for embodying an avatar and 3D Audio capturing.

The blue *Artificial Intelligence components* deal with artificial intelligence techniques for perception, cognition and animation. They provide higher level forms of intelligence to bridge the semantic gap between the naturalistic (camera captures) and synthetic (computer animated) objects. Desired functionalities include emotion understanding and conversational capabilities of avatars.

The green components deal with networking, streaming and compression functions. As the configuration presented in this system includes 3D reconstruction, capturing, 3D Rendering, AI and social network functions, specific network support is required. Therefore, the presented architecture presents a fertile ground for research and experimentation with network protocols to support the requirements of this application. This thesis will develop the appropriate network and codec support for this architecture to tackle research challenges arising from this configuration.

The pink *rendering components* are components for 3D rendering related functions. They include spatial audio composition (mixing of multiple audio streams for efficient delivery to end users). Scene structuring and navigation (handling the composition of all objects in the 3D Room and mapping them to a single space). At the client sides, different renderers are implemented for different types of 3D data. In this architecture these renderers can operate in parallel and their output are composited, rendering multiple objects in a single scene.

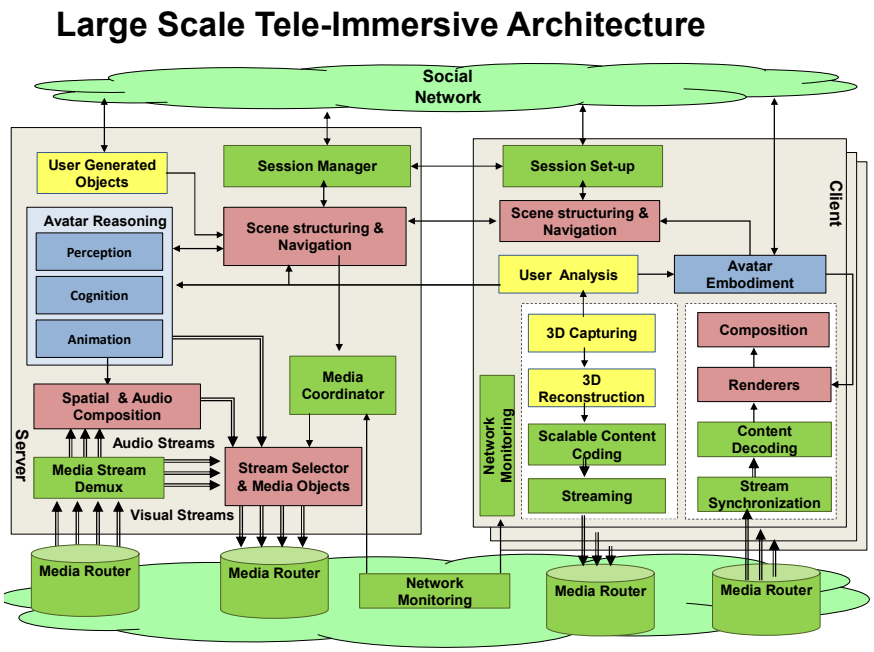

Figure 2 Large Scale Tele-immersive System Architecture

1.4 Research Questions

The architecture as shown in Figure 2 will enable 3D immersive mixed reality. In such a scenario naturalistic and synthetic contents will be mixed in a virtual world, resulting in a novel type of distributed shared experience beyond gaming and video conferencing. To support such applications from the network side, the main research question addressed in this thesis is the following:

Main Research Question: How can we support Real-Time (<300ms) end-to-end media streaming for Mixed Reality and 3D Tele-immersion be supported in the current Internet with state of art compression rates ?

This thesis addresses this question using an experimental integrated multimedia systems approach. It will build and test prototype systems that include components from capture up to rendering, thus integrating components for networked transmission and

3D compression in complete systems. The advantage of such a system centric approach is that it can shed light on integration issues and optimization strategies between components that have been previously overlooked. In addition, the prototypes can be tested with real users, resulting in explorative human centric studies on the benefits of different technical implementations. In addition, the systems and human centric approach can lead to preliminary requirements of importance for the development of the technology in academia and industry. To this aim the work in this thesis has contributed to updates in the requirements of media technology standards in MPEG [8] and JPEG PLENO [9]. The main research question is split in 5 questions that are addressed throughout this thesis.

Research Question 1: *What is the state of the art to support tele-immersive mixed reality, and how does it relate to other 3D media applications?*

The research question aims to explore related work to find out what kind of mixed reality and 3D communications systems have been developed in the past and which technologies are available for the next generation of systems. It reveals that previous work does not target the architecture that this thesis envisions and that this architecture poses many challenges. Further, we present a generic pipeline for development and study of multimedia systems and standardisation of related data types (compression formats, raw digital capture formats and network ready bit streams). This pipeline forms a reference model for standardisation of different audio visual datatypes in MPEG, and was partially based on a contribution of the author during this thesis work (see section 0).

Research Question 2: How can we achieve Real-time (below 300 milliseconds on commodity hardware) compression and transmission of highly realistic reconstructed 3D humans based on meshes with State of Art Compression rates (i.e. MPEG-4)?

Our second research question deals with the real-time streaming of highly detailed reconstructed 3D humans from multiple range sensors as 3D meshes. This is highly challenging due to the number of points (over 300,000) in the mesh and the rate of capturing (in the order of 10 to 12 frames per second) and the real-time transmission requirements. A closer look at the state of the art in mesh codecs shows that their

complexity increases linearly with the number of vertices and that especially real-time encoding is challenging and ill-considered in previous work. In addition, compressed 3D data is prone to data losses. Therefore, the aim of this research question is also to investigate a reliable transmission protocol that also works in real-time.

Research Question 3: How can we achieve highly adaptive (lossy) real-time compression and transmission of highly realistic 3D replicants based on meshes that can be used for many different bit-rates and level of detail (i.e. a geometry coding with a large range of bit-rates)?

To support real-time conferencing in the real Internet where bandwidth fluctuations happen, a codec that can support many different bit rates is beneficial. In addition, as known from computer graphics, 3D objects can often be rendered at lower levels of detail. This is especially the case if objects are further away in the scene. Based on these motivations, this thesis explores the design of a mesh codec that runs in real time and is highly adaptive. This codec represents a lossy design, i.e. not all input vertices/triangles are present in the decoded version. Such a codec results in a lower bit rate and a lower quality decoded output. This is useful for multi user/ multi-site teleconferencing using 3D mesh data and for less relevant 3D reconstructions that can be rendered at a lower quality.

Research Question 4: How do we design improved Real-Time Compression of 3D Time varying Point Clouds with improved lossy colour coding and inter-prediction with a negligible perceptual distortion compared to the current state of the art in practical real-time point cloud compression?

Point clouds are simpler to acquire than meshes and introduce less overhead as they do not store the connectivity information. While real-time point cloud compression has been studied in the past, some aspects relevant to mixed reality systems have been poorly addressed. Therefore this thesis investigates three aspects of real-time point cloud compression. First it studies real-time and lossy coding of colour attributes. Second, it will explore lossy inter predictive coding of 3D tele-immersive point clouds. Last, it will investigate quality assessment methodology based on an objective metric and subjective evaluation in a mixed reality system. The thesis combines the research results to study the complete point cloud compression framework for

mixed reality and tele-immersion and assess the performance both objectively and subjectively.

Research Question 5: how can we design a 3D media streaming engine for heterogeneous 3D tele-immersive communications that is compatible with the current internet infrastructure (i.e. existing transport and signalling protocols)*?*

This thesis will investigate the support of 3D Media streaming in mixed reality with various 3D capture and render modules. It will investigate the type of transmission schemes needed such as peer-to-peer, publish-subscribe, application layer multicast. It will develop a unified API for the external 3D modules to use the network streaming facilities. Subsequently, it investigates session management for presence and stream setup. Lastly, support for media synchronization is added and the integration into the mixed reality system is completed. The thesis will evaluate the streaming system in a realistic mixed reality system with multiple sites.

1.5 Contributions

This thesis presents the following contributions to the field of multimedia systems:

It presents work on a ***novel architecture and framework for 3D Tele-immersive mixed reality*** for use in the Internet. This framework includes advanced modules for 3D capturing, rendering and Artificial intelligence in the context of an immersive virtual room linked to the social network. This has not been considered in previous works. This is of particular relevance for enhancement of social networks for shared distributed experiences. These insights will benefit large technology providers interested in providing such services such as Facebook (Oculus), Microsoft Skype, Linkedin, Google Hangout, World of Warcraft etc.

It presents ***novel approaches for Geometry Coding suitable for 3D Tele-immersive mixed Reality***. Contrary to previous attempts in 3D graphics literature, it develops encoder/decoder frameworks that work in real time on commodity hardware. This is critical for enabling real-time end-to-end communications for 3D tele-immersive Mixed Reality. It starts with approaches that make use of the capturing algorithm/procedure that can benefit the coder, optimizing the end-to-end pipelines. Later, it develops complete and generic highly adaptive real-time mesh and point

cloud codecs. The generic geometry driven mesh codec has been integrated in the tele-immersive framework for multi-site streaming. Further, this thesis introduces a time varying point cloud codec implementation that follows a hybrid architecture and includes techniques for inter-frame coding and lossy colour coding. The point cloud codec implementation has been contributed to MPEG and currently serves as a base implementation for development of a standard for point cloud compression in MPEG.

It presents ***Approaches for Network Streaming and Signalling for 3D tele-immersive and mixed reality***. It proposes a signalling and session framework based on XMPP protocol to setup sessions between heterogeneous clients as a control plane. For 3D geometric data transmission experimented with the TCP and UDP protocols using LT codes are performed. Further, the end-to-end multithreaded pipeline for 3D geometry transmission is enhanced using a last in first out policy (LIFO) to reduce end-to-end delay. The overall streaming framework includes a generic API for stream creation and a virtual clock system that can enable media synchronization (both inter-stream and inter-sender). Last, the framework includes a fast messaging system for publish and subscribe based transport. This is useful for command and other animation messages. This framework provides core network support for a larger 3D Tele-immersive mixed reality system developed in the context of the EU project Reverie [10] and provides hints on how such network support can be provided in future tele-immersive systems.

It presents **Systems integration into large scale prototypes and preliminary user centric evaluations.** All components have been integrated in a larger 3D Tele-immersive mixed reality framework. The integration of different components for 3D capture, data transmission and rendering provides some insights of potential bottleneck issues and many practical issues. The full systems integration has enabled user testing in real world scenarios. These tests all hint towards the benefits of highly realistic 3D representations and 3D tele-immersive communication in collaborative tasks and social interaction tasks.

It contributed to the **development of requirements, benchmarking tools and codec platform implementation for future video and image coding standards.** The systems integration and codec development presented in this thesis unlocked novel

requirements for future image and video coding. Over the course of 3 years, it has contributed to the international standards for media technology (MPEG) [8]. This has resulted in updates to support time varying meshes and point clouds in the MPEG requirements [11]. In addition documents defining the requirements of 3D tele-immersive coding have been ratified in MPEG [12]. In addition, the methodology (including the quality metrics and benchmarking evaluation platform developed in this work) for point cloud compression have all been adopted by MPEG in the activity on point cloud compression [13]. The Point cloud compression and evaluation platform is used for the further development of the Point Cloud Compression in MPEG in the coming years, serving as a base exploration software. Last, during this work, the author has contributed to the draft version of the JPEG PLENO [9] [14] requirements, an emerging 3D Visual standard that will target compression of naturalistic point clouds, light fields and holographic images. These contributions focussed on the use cases in Mixed Reality 3D Tele-immersive systems area.

1.6 Thesis Outline

Chapter 2 presents related work and technical background based on the current advances in the Internet, 3D capture, rendering and social networks. Further it presents a generic 3D audio visual capture reference model for multimedia systems and data format standardisation. Last, an overview of previous work on 3D tele-immersive systems is presented. This chapter answers research sub question 1.

Chapter 3 presents a system centric optimized real-time streaming component for reconstructed 3D mesh data (capturing based on [15]). It includes a compression and transmission component. It exploits both regularities in the connectivity and in the geometric blocks and is computationally simple and easy to parallelize. In addition, the prototype includes a novel UDP based network transmission scheme based on a linear rateless code. It compares the end-to-end performance of the prototype with the MPEG-4 mesh codec and transmission based on TCP. The prototype reports reduced end-to-end delay resulting in real-time performance. A major conclusion of this work is that traditional schemes for mesh coding and transmission are not designed with the requirements of 3D tele-immersion in mind. This chapter further partially answers sub research question 2. The work in this chapter is based on the following two publications.

Mekuria, R.N. *Sanna, M. Asioli,S. Izquierdo, E. Bulterman, D.C.A. and Cesar, P. A 3D Tele-Immersion System Based on Live Captured Mesh Geometry. Proceedings of the 4th ACM Conference on Multimedia Systems (MMSys 2013), Oslo, Norway, February 27- march 1 2013 (focussing on the network-compression pipeline integration) (Best Paper Award in special session on 3D in Multimedia)*

Mekuria, R.N. *Alexiadis, D. Daras, P. and Cesar, P. "Real-time Encoding of Live Reconstructed Mesh Sequences for 3D Tele-immersion." Proceedings of the International Conference on Multimedia and Expo. International Workshop on Hot Topics in 3D (Hot3D) San Jose, CA, July 19 2013 (focussing on the capture-compression integration)*

Chapter 4 presents a codec based on connectivity driven mesh coding and an optimized streaming pipeline that integrates improved rendering. The codec bears similarity to the codec defined in the MPEG standard for mesh compression [16] as it uses the connectivity to efficiently code the geometric positions. By exploiting regularities in the connectivity using differential and entropy coding based on the deflate/inflate algorithm and repetitive patterns, the connectivity coder is made highly efficient. The solution is fast and more generically applicable compared to the codec developed in Chapter 3. Subsequent to the connectivity coding, late (staged) quantization of differentials between connected vertices to code the geometry is performed. This allows the codec to skip explicit variable length entropy encoding. Next, the streaming pipeline is optimized using a last in first out approach (LIFO) for inter thread exchange of frames. The overall system is tested in scenarios representing typical moderately managed networks such a VPN Network. The overall optimization in coding and transmission enables real-time end-to-end communication of the highly reconstructed 3D human combined with highly realistic rendering using global illumination. This chapter answers research question 2 in a more generic way. The chapter is based on the following publications.

Mekuria R.N., *Cesar P., Bulterman D.C.A. "Low Complexity Connectivity Driven Dynamic Geometry Compression for 3D Tele-Immersion" in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (codec) (IEEE ICASSP 2014), Florence, Italy, May 4-9, 2014 (codec implementation)*

Mekuria R.N., *Sanna M., Izquierdo E., Bulterman D.C.A., Cesar P. "Enabling 3D Tele-Immersion with Live Reconstructed Mesh Geometry with Fast Mesh Compression and Linear Rate less Coding" IEEE Transactions on Multimedia 2014 Volume 16 issue 7 pp. 1809 – 1820 (codec, integration, pipeline optimization)*

Chapter 5 presents a highly adaptive geometry driven real-time mesh codec for use in 3D tele-immersive applications. Contrary to the codecs developed in chapter 3 and 4 and the codec in MPEG-4 [16], this codec enables fine grained bit-rate and quality control. It enables elimination of vertices (mesh simplification), which for 3D geometric data is a better way to control quality and bit-rate compared to vertex quantization (this is performed in MPEG codecs such as MPEG-4 TFAN). The mesh geometry is organized and simplified in an octree voxel grid. Subsequently the simplified connectivity is computed based on the octree cells and the original connectivity. Then, the new connectivity representation is converted to a novel representation consisting of the first index, and two vector integer offsets in the octree grid. Subsequently this chapter develops an approach for vector quantization to store the two offsets. The first index is coded in a run length fashion, the second two based on this vector quantization scheme. This approach achieves a compact representation of the connectivity, exploiting the grid organization of the octree. The octree structure is then compressed using serialization and a range coder. The key benefit of this codec is that it can be used for multi-site streaming and rendering less relevant objects such as those at distance at a lower level of detail. This chapter addresses research question 3 and was published in the following works.

Mekuria R.N. *Cesar P. "A Basic Geometry Driven Mesh Coding Scheme with Surface Simplification for 3DTI" IEEE Communication Society: Multimedia Communications Technical Committee E-letter, May 2014 (Codec implementation)*

Mekuria R.N. *, Cesar P, Doumanis I, and Frisiello A., "Objective and Subjective Quality Assessment of Geometry Compression of Reconstructed 3D Humans in a 3D virtual room," in Proceedings of the Applications of Digital Image Processing XXXVIII Conference, (part of the SPIE Optical Engineering + Applications, part of SPIE Optics + Photonics symposium) , San Diego, USA, August 9-13, 2015 (subjective evaluation, system integration)*

Chapter 6 presents a novel design for real-time point cloud compression for 3D tele immersive video. This chapter deals with some specific challenges for compression of point clouds for 3D tele immersive video. It starts from the classical octree based approach for point cloud compression. Then, it introduce schemes for improved lossy real time colours coding by mapping to a JPEG image grid. Subsequently, it introduces a temporal inter-prediction scheme between frames by computing motion vector translations between occupied macro voxels (blocks of octree leafs). It efficiently stores 3D rotation as a quaternion, that is compressed using a known quaternion compression scheme and quantizes the translations using 16 bits. Points that could not be predicted are coded in intra fashion (using an octree serialization and compression). This is followed by objective rate distortion evaluation followed by a subjective user study with 20 users in a realistic mixed reality system. In addition a parallelized implementation using OpenMP is provided. The results show that the inter-prediction can save up to 30% in bit rate, while the colours coding scheme results in highly reduced bit-rates (up to 90 %), without being noticeable by users. The chapter is based on the following publications.

Mekuria R.N. *Blom K., Cesar P. "Design, Implementation and Evaluation of a Point Cloud Codec for 3D Tele-immersive Video" IEEE Transactions on Circuits and Systems for Video Technology (accepted for publication, to appear sept. 2016)*

Mekuria R.N *"Point Cloud Compression" proceedings of inaugural workshop on JPEG PLENO Workshop in conjunction with MPEG/JPEG meeting in Warsaw, June 23, 2015*

Mekuria R.N. *Cesar P. MP3DG-PCC, Open Source Software Framework for Implementation and Evaluation of Point Cloud Compression.in Proceedings of the 2016 ACM on Multimedia Conference (MM '16). ACM,*

In addition, the implementation has been accepted as a first implementation for point cloud compression MPEG-4 part 16 (AFX) in the MPEG standardisation software repository [13]:

Chapter 7 presents a content-aware 3D streaming engine and the integration in the overall framework. It integrates the components developed in previous chapters, Further, the system provides the API to the overall mixed reality system for multi-site and multi-stream communication. This component supports creation of different types of media streams (visual, audio etc.), lightweight messaging (publish and subscribe) and control messaging. It uses an XMPP server for presence and stream setup, using an extension of the XMPP XEP-30 disco protocol. This presence protocol aims to demonstrate how, based on available modules, social 3D tele presence in a virtual world can be supported. The framework provides a synchronization mechanism based on a virtual clock. Synchronization schemes based on buffering have been implemented in the rendering modules. The streaming framework has been evaluated for multi-site streaming in the larger Reverie tele-immersive system. The synchronization subsystem achieves approximate inter-stream and inter-sender synchronization of 3D audio and 3D geometry streams coming from different sites. The work has been presented in the following scientific publication.

Mekuria R.N. *Frisiello A., Pasin., M, Cesar P.S. "Network Support for Social 3-D Immersive Tele-Presence with Highly Realistic Natural and Synthetic Avatar Users" in ACM workshop on Massive Multi User Virtual Environments MMVE'15, Portland USA, in conjunction with ACM MMSys'15*

Chapter 2 Related Work

This chapter presents the related work. First, it provides an overview of the technical background in 3D capturing, rendering and networking. Then it will provide some related work on 3D Tele-immersive systems development. This chapter addresses research question 1: *What is the state of the art to support tele-immersive mixed reality, and how does it relate to other 3D media applications?*

2.1 Technological Background and Motivation

2.1.1 3D Scene Capture

3D cameras detect not only pixel colours, but also the depth information of the scene. The depth information is important as it can be used for rendering arbitrary viewpoints via depth image based rendering (DIBR) which, amongst others can be used for stereoscopic rendering. For mixed reality and 3D tele-immersive communication the depth information can also be used to segment 3D objects of interest. Examples of such objects of interest are the human or participant or buildings or landscapes, or other objects such as an automotive vehicle. This segmentation often results in 3D Meshes or Point Clouds which are vertex positions and triangles in the 3D space. The segmentation can work especially well when multiple depth cameras are recording from different angles and are calibrated (both internally: depth and colour and externally between cameras). This section briefly summarizes technologies for sensing depth data, followed by camera configurations for 3D reconstruction. The chapter also summarizes an example 5 camera setup to reconstruct 3D Meshes and Point Clouds based on [15].

Depth sensing from single passive camera: 3D shape can be obtained from sequences of single camera video data. These techniques are often described as *motion from X techniques* that use shading, texture, focus or motion (for X) to obtain the 3D structure of the scene. Even though there is no complete and fully reliable solution to obtain 3D structure from a single camera stream, techniques based on structure from motion tend to be the most promising [17] in this class of techniques.

Depth sensing with multiple passive cameras often implies the usage of two (or more) configured cameras and solving some under constrained correspondence problem. A common method is *stereo matching*, that bears similarity with how the human visual system senses depth with two eyes based on binocular cues. Based on the two input images and known camera parameters and positions, disparity can be calculated between corresponding points in both scenes. This disparity information can then be used to estimate the depth at the corresponding points in the scene. However finding such matching points with a computer does not always work well. In case of occlusions and repeating textures the problem of finding corresponding points in the scene becomes ambiguous [18] and depth estimations can fail.

Depth sensing with active cameras is a more direct and straight forward way of acquiring depth information from a scene without additional computational efforts. Microsoft introduced the popular Kinect 1 sensor which is based on *structured light*. It is based on transmitting an infrared dot pattern with an infrared projector, and receiving it with an infrared Receiver. Triangulation of the received and original patterns can be used to compute the depth information. The principle is shown in Figure 3 where Zo is the distance to the object, Zr the distance to the reference plane (i.e. maximum depth), the distance between the IR receiver and projector and d the disparity between the two triangulated patterns and f the focal length. Depth can then be calculated based on formula (2.1).

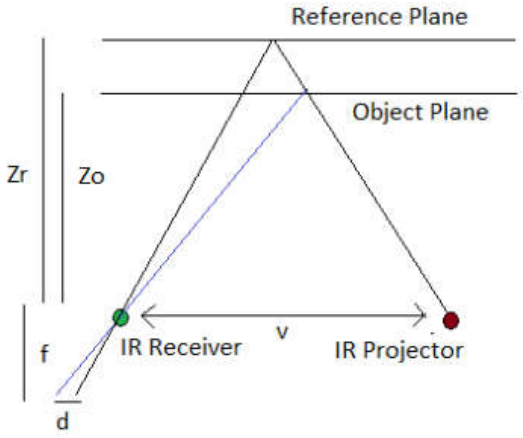

Figure 3 Triangulation of infrared pattern based on structured light [18]

$$Zo = \frac{Zr}{1+\frac{Zr}{fv}d} \quad (2.1.)$$

Alternatively, the Microsoft Kinect for Windows v2 is based on Time of Flight (ToF) principle, which is also used in LiDaR (Laser Imaging Detection and ranging using lasers instead of infrared light). The ToF depth sensing principle is based on measuring the distance the light has travelled. This is done by modulating the light and measuring the phase shift introduced at the receiver compared to the original signal. If the light is modulated at frequency f_{mod} the depth of the object Zo can be computed by formula 2.2. In this equation c is the speed of light, f_{mod} the modulation frequency and φ_d the measured phase shift in radians. LiDars use lasers instead of light at wavelengths near the infrared spectrum and have much larger range and resolution. These active techniques provide easier and more reliable ways to sense 3D depth information with less computational overhead compared to passive methods.

$$Zo = \frac{c}{4f_{mod}}\frac{\varphi_d}{2\pi} \quad (2.2)$$

Figure 4 Microsoft Kinect, v1 (left, based on structured light) and v2 (right, based on Time of Flight ToF) and SoftKinetic (ToF)

An alternative and generic way to capture a naturalistic 3D scene is based on the plenoptic representation. Plenoptic cameras try to capture the light fields produced by the 3D scene by capturing for each (x,y,z) the luminance radiated in different directions (θ, φ) over any range of wavelengths λ at every time t. This results in a 7 dimensional function for luminance (eq. 2.3.). Sparse Capturing of these light fields is usually done with many cameras/pinholes that are spatially co-located. Storing this type of data is highly challenging as the 7 dimensions introduce large amounts of data.

$$l(V_x, V_y, V_z, \theta, \varphi, \lambda, t) \quad (2.3)$$

For more information on handling this type of data we refer to [19]. The rest of the thesis will look at object segmentation from colour plus depth data, instead of this type of plenoptic data. Segmented objects and a synthetic scene will allow us to implicitly and artificially generate approximate instances of $l(V_x, V_y, V_z, \theta, \varphi, \lambda, t)$ by applying rendering techniques from computer graphics. This enables using the more sparse geometry representation instead of a heavy plenoptic representation. In addition, it introduces higher flexibility for our mixed reality system and enables composite rendering with synthetic 3D graphics objects more easily.

To segment 3D objects from colour data plus depth data (either from passive or active depth sensing), multi camera setups combined with efficiently implemented 3D reconstruction algorithms can be used. Reconstruction algorithms generally remove background and other irrelevant data and stitch the point clouds from each view together and compute a single 3D Mesh or Point Cloud. Several attempts have been made with Microsoft Kinect technology [15], [20], [21] and in the near future further advances are expected. In the figures below we show some of the full 3D meshes obtained in real-time from multiple calibrated Microsoft Kinect v1 of time varying human subjects. On the left side techniques have been based on Zippering based on [15] while in right side we show the results obtained with Poisson 3D reconstruction based on [22]. The observed artefacts are due to infrared interferences between Kinect and stitching errors. With higher resolution sensors like Microsoft Kinect v2 and SoftKinetic it is expected that better 3D reconstructions will be possible. However, experimental rendering of these meshes in real time has shown that they give a photorealistic impression despite these artefacts. In addition, technologies for 3D point cloud capture are now also becoming available on mobile handheld devices [3] [4].